DOCUMENT RESUME

ED 222 501 TM 820 542

AUTHOR Cuttance, Peter F.
TITLE Covariance Structure Modelling of Reliability and Differential Response in Educational Survey Data.
PUB DATE Mar 82
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Covariance; Correlation; Educational Research; Error of Measurement; *Item Analysis; Mathematical Models; Regression (Statistics); *Response Style (Tests); *Surveys; Testing Problems; Test Items; *Test Reliability
IDENTIFIERS *Covariance Structure Models

ABSTRACT

Covariance structure modelling is applied to the problem of estimating reliability and measurement error in survey data. To provide a basis for grouping certain question or variable types (data from questions), a simple typology based on the formal characteristics of the questions is outlined. From this classification, models for the different question types are used to estimate reliability and method influences on the observed responses. Differential reliability across questions within each question type is investigated in terms of a subsidiary hypothesis that the degree of reliability varies systematically according to the substantive context of a question. Estimates of measurement error are employed to disattenuate a correlation matrix and thereby demonstrate the differences obtained from correlation and regression analyses when these issues are ignored. (Author/CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED222501

COVARIANCE STRUCTURE MODELLING OF RELIABILITY AND DIFFERENTIAL RESPONSE IN EDUCATIONAL SURVEY DATA

Paper prepared for presentation at American Educational Research Association Annual Meeting, Division D, New York, March 1982.

Peter F Cuttance
Centre for Educational Sociology
University of Edinburgh
Edinburgh EH8 9LN

Reference No. 8206
09/02/82

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. F. Cuttance

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 542

Covariance Structure Modelling of Reliability and Differential Response in Educational Survey Data

Introduction

In Part I of this paper I apply covariance structure modelling to the problem of estimating reliability and measurement error in survey data and, in Part II I demonstrate the consequences of assuming that the data are not fallible when in fact they are. In order to provide a basis for grouping certain types of question or variable (ie data from questions) a simple typology based on the formal characteristics of the questions is outlined. From this classification, models for the different question types are used to estimate reliability and method influences on the observed responses. Differential reliability across questions within each question type is then investigated in terms of a subsidiary hypothesis that the degree of reliability varies systematically according to the substantive context of a question. Estimates of measurement error are employed to disattenuate a correlation matrix and thereby demonstrate the differences obtained from correlation and regression analyses when these issues are ignored.

Part I: Modelling the Reliability of Response in Survey Data

1.1 Theoretical Orientation

Measures of Reliability

Lord and Novick (1968) have suggested that any of the following three quantities might be used as a coefficient of reliability:

(1) the correlation between two parallel measures,

(2) the degree of variation in the latent variable relative to the variation in the observed variable or,

(3) the squared correlation between the observed variable and the latent variable.

Definitions (2) and (3) rely on notions of an observed (manifest) variable being a measure of some underlying theoretical or latent variable. For estimation purposes all three definitions require there to be at least two (repeated) observations on the same manifest variable or, alternatively, at least one observation on two manifest variables measuring the same theoretical variable. The latent variable in definitions (2) and (3) is defined as the theoretical variable or true score of interest. The variance in the observations on the manifest variables is then defined to be the sum of components due to the theoretical (latent) variable plus that due to the error of measurement. Diagramatically this can be shown as in Figure 1 where X_{11} and X_{12} represent manifest variables measuring the same theoretical variable (ξ) and δ_1, δ_2 are random errors of measurement.

[Figure 1 about here]

Question Types

In educational survey data it is normal to find a substantial range of variation in reliability across variables. Broadly speaking, two approaches to classifying variables into categories with relatively homogenous degrees of reliability have been suggested. First, researchers involved in the analysis of data have suggested grouping variables on the basis of their substantive content. Second, survey methodologists have suggested that the formal characteristics of variables may provide a suitable classification. In this paper I put forward a typology based on this second line of thought and then explore the variation in reliability for questions of different substantive content in this context. From the literature on survey response/non-response I have derived a typology of question types which provides a basis for classifying the data from surveys conducted on school leavers in Scotland.

Following Kalton and Schumann (1981) the first distinction I make between the types of information sought from different forms of question is that referred to as 'factual' vs 'non-factual' information. A factual question will be defined as one for which,in principle, the correct response could also be determined from at least one other source in addition to that sought on the occasion stipulated.An exception to this definition is factual information about an individual's private life of the mind. In contrast, non-factual questions cannot be validated by reference to sources external to that sought. Validation is only possible for non-factual questions by assessment of the consistency of a set of data for different questions which all aim to measure the same underlying theoretical construct, ie through the assessment of construct validity.

In addition to the factual/non-factual distinction made above, Kalton and Schumann posit several other distinctions between question types which may have a bearing on the reliability of a question.The following are those which are relevant to the questions used to obtain the data in the current study:

(a) Reason questions - questions which ask the respondent to indicate their reason for a particular,named,action which occured at some specified time in the past.

(b) Judgement (or perceptual) questions - these are questions which contain a factual component overlaid with the respondents evaluation of that fact.Since the correlation between respondent's evaluations of externally defined facts and the external evaluations of those facts may be quite low,this is not considered to be a particularly valid means of obtaining factual information.Used as a means of obtaining information of a respondents' perception of particular factual information,this method is similar to asking respondents for their opinion on those facts.

(c) Recall questions - these are questions which ask the respondent to recall information relating to a specific period in the past,and may be of either a factual or non-factual character.

(d) A further distinction which has been alluded to in the literature is that which differentiates information about oneself from that about others.

(e) When responses to questions are interpreted as being measures of particular variables, the distinctions in (a)-(d) above carry over to the variable interpretations. Variables which are based upon more than one question may, however, conflate these distinctions. For the purposes of the current study such variables will be referred to as 'derived' variables. The classification of derived variables within the above typology depends upon the classifications attached to their constituent variables.

Figure 2 indicates one way in which the above typology could be constructed so as to represent an hierarchical structure of question domains.

[Figure 2 about here]

1.2 Questionnaires and Data

The data analysed in the present study was collected as part of the 1977 Survey of School Leavers in Scotland(Raffe,Lamb et al , 1977).The questionnaires employed in collecting the two waves of data used here were those denoted 'version R ' and 'version Z' of the set employed in the overall survey.Version Z was an updated but minimalliy modified facsimile of a questionnaire which had been administered in a similar survey conducted in 1972 (cf. Jones and McPherson, 1972 for details of the proceedures employed) while version R was a newly designed questionnaire for the 1977 survey.The population to whom these questionnaires were administered in 1977 was the group of school leavers who had attempted at least one Higher examination prior to leaving school during the 1976 school year.

The data were collected under a 2 x 2 factorial design in which a selected sample of leavers were sent a questionnaire in March 1977.From

the group who responded to this first questionnaire a subsample was selected and sent the second questionnaire in May 1977.The second factor in the design was fixed by the order of the administration of the two questionnaires. Approximately half of the leavers received version R as their first questionnaire and half received version Z.In the second wave of questionnaire administration, the alternative questionnaire to that sent on the first occasion was administered.The two data sets based on these groups of leavers are referred to according to the order in which the leavers received the questionnaires,thus they are denoted the 'RZ' and 'ZR' administrations, or data sets.

The response rate for version R when administered first, was in the order of 80 per cent for the total sample who received that version for the main survey. Comparable figures for the other three administrations of the questionnaires are not available. The numbers of respondents for whom both questionnaires were treated as valid returns on both occasions is 225 and 267 for the RZ and ZR data sets,respectively.

The questionnaires contained only a small group of questions which attempted to obtain the same information in both cases. The current study of reliability focuses on this subset of questions as measures of the same theoretical variable on both occasions. We refer to the observed variables which measure the same assumed underlying variable across occasions as replicated variable pairs. Table 1 lists the variables derived from these questions and indicates points of dissimilarity within the variable pairs. The variables which employed different response categories in the two questionniares have been recoded where possible to make the measurement scales in the data equivalent for both members of a variable pair.In some cases,however,such recoding was not possible(See Appendix A for details).

[Table 1 about here]

The number of valid response cases in each data set varies across variables,but to only a minor extent.The data employed in the various analyses is ,however,based upon listwise deletion of missing cases,thus

the number of cases for any given analysis includes only those individuals for whom data is available on all variables in the particular data subset. This procedure results in data sets for the various analyses which are based on more than 190 cases in each instance, with the exception of that for the WHYFE variables. These latter variables are based upon questions which were applicable only to those who were in further education at the time of the survey, and result in data sets of 155 and 167 cases for RZ and ZR , respectively.

1.3 Model: Estimates of Reliability and Response Error

Although the simple model of Figure 1 implied that the only errors of measurement are those due to random disturbances (δ_1, δ_2), this model will generally not fully describe the situation. A more appropriate model in most cases will be that shown in Figure 3. The difference between Figures 1 and 3 is that the error terms are now respecified as partitioned residual terms. The total variation in the error term is redefined to consist of the sum of a systematic or 'specific' part plus variation of a random nature. The specific variation may be due to method or other occasion specific influences on the observed score. If we have just the two measures X_{11}, X_{12} as in Figure 3, it is not possible to estimate the amount of variation due to the specific and random components of the residual term separately. This provides the raison d'etre for the typology of question types put forward above. For, the effect of such a classification is to provide the basis for locating specific components of the residual variance as common variation across the residual terms in the model. It is hypothesised that questions of a given type will contain common (specific) variation due to method or other occasion specific influences.

[Figure 3 about here]

The general model based on this formulation is that shown in Figure 4. I will refer to this as the multimethod-multiquestion (MMMQ) model. This designation alludes to its similarity in both form and interpretation to that known as the multitrait-multimethod (MTMM) model

found in social psychology (Campbell and Fiske, 1959; Joreskog, 1971). Since the model has more than one indicator for the latent variable measuring method effects, it is possible to impose restrictions which over-identify the model. Hence, it is possible to estimate and test for the components of variance due to method effects and other residual (measurement) error. We note that one form of restriction would be to specify that the measures are 'parallel'. This implies that $\lambda_{11}=\lambda_{21}$, $\lambda_{12}=\lambda_{23}$ and $\delta_{11}=\delta_{12}$. However, the model can be identified by fewer restrictions than this, thus it is possible to test the assumption of parallelism, and also to estimate less restrictive models. In general, we expect the assumption of parallel measures to be too strong for most of the data obtained in educational surveys, hence estimates based on definition (1) in section 1.1 will be biased to an unknown degree (cf. Ecob & Goldstein, 1981).

[Figure 4 about here]

The analysis of the data with the model presented in Figure 3 was performed with the program LISREL IV (Joreskog and Sorbom, 1978). The procedure followed was to estimate and test the fit of separate models for the different question types. First, I tested for the form of the measurement model for the replicated variables in each model and, second, I tested for differences in method effects across and within occasions.

Five hypotheses relating to the form of the measurement model were tested:

Hypothesis H1.0: A measurement model specified as congeneric provides an acceptable fit to the data and is preferrable to the null hypothesis. The null hypothesis is defined as the unconstrained (saturated) model of the relationships between the latent and observed variables.

Hypothesis H2.0: A measurement model specified as tau-equivalent fits the data and is preferred to the null model.

Hypothesis H2.1: A model specified as tau-equivalent provides a more acceptable fit to the data than one specified as congeneric.

Hypothesis H3.0: A measurement model specified as parallel fits the data and is preferred to the null model.

Hypothesis H3.1: A model specified as a parallel measures model provides a more acceptable fit to the data than one specified as tau-equivalent.

Similarity in the behaviour of the method factor across occasions was interpreted through the correlation between method factors. Further, a series of hypotheses relating to the behaviour of the method effects within and between occasions were tested for the individual manifest variables in the models. From this hypothesis testing procedure I determined the specific form of the model for the group of variables representing a particular question type. The model chosen as giving the most parsimoneous acceptable fit to the data is referred to as the 'preferred model'. The rather lengthy statistical details of this model testing sequence are not reported here but can be found in Cuttance (1981b). A description of models for each of the three question types investigated is given in Table 2.

[Table 2 about here]

The models resulting from the foregoing analyses were employed to estimate the reliabilities of the various questions investigated in this study. Since the LISREL IV program requires that the data correlation matrix be positive definite, and, hence, that it does not contain redundant rows or columns, it is not possible to estimate models containing variables for which the correlation between observations on the two occasions is unity. Formally, a correlation of unity implies that it is not possible to distinguish between the influence of the true score and method effects on the observed score. Further, it implies that there is no random error of measurement and that any method effects which are present, covary perfectly with the 'true score' for the variable. Thus, the product moment correlation coefficient itself provides a suitable measure of reliability, under definition (1) above,

for these variables.

As implied by the model in Figure 4 the variance components which measure the reliable influences on the observed score are due to both method and true score effects. Both components may contribute to the reliability of a measure but it is clear that only the true score component is a 'valid' measure of the theoretical variable of interest. Thus the total reliability of the measure is due to a valid, plus an extraneous or 'invalid' (method) component. Any variation across occasions which is due to change in the underlying theoretical variable will reduce the component due to the true score effect and increase the amount of variance accounted for by the method and error terms.

Table 3 presents summary statistics, estimates of reliability and variance components for the variables. Blanks within the table indicate variables for which it is not possible to estimate MMMQ models; due to correlations close to unity.

[Table 3 about here]

In this section I wish to consider the findings from the analyses in the light of differential relibility within each question type category, but I will first provide a brief summary of the findings of methodological significance related to question types/variables reported elsewhere (Cuttance, 1981b).

The variables reported in Table 3 are listed by the question type classification outlined in section 1.1 above. Questions of a factual nature are more reliable than those of a non-factual nature and, within this class, those relating to questions about oneself are more reliable than those about others. Further, questions of a factual character requiring recall of past events are slightly less reliable than those asked about events of the present. All of the questions of a factual character about oneself have a reliability close to unity. Thus they could be treated as if they were perfectly reliable for the purpose of multivariate analyses but those relating to 'others' should be used only in conjunction with their reliability estimates.

Questions of a non-factual nature should also be used only in conjunction with estimates of their reliability. This is true of both sub-classes of question investigated: reason and perceptual questions. Again there is a suggestion in the findings that recall questions in this class may be less reliable also, but that the main problem in the use of such questions may derive from changes in the underlying 'true scores' across occasions.

Estimates of reliability based on the correlation between the observed scores rely on an assumption that the only systematic component of variation in the observed score is that due to the true score. Thus it is most defensible to use such estimates when the correlation between measures is close to unity, since this also implies that there are no systematic method influences which are orthogonal to the 'true score' influences on the observed responses. However, in cases where there are non-true score influences of a systematic nature, then estimates based on correlations between measures are likely to be biased. The findings confirm that this bias may be either positive or negative, hence it is necessary to give consideration to models which can account for such extraneous influences, in the design of a survey. By the selection of judicious design strategies it is possible to provide a structure for the data which allows for estimation of the different components of reliability in the observed scores.

In discussing the findings with respect to differential reliability within question types I will use the estimates provided by the variance components approach of the MMMQ model. The variables of most interest in this context are those which one cannot reasonably use without concommittent adjustment for their degree of reliability. I consider the content of the questions within each question type as a means to investigating possible links between substantive content and the degree of reliability.

As noted above, questions of a factual nature about oneself were found to have a reliability close to unity. This means that their reliability is invariant to the substantive content of their inquiry. In this investigation this finding relates to questions covering formal

biographical characteristics of age, sex, when left school and, the number of brothers and sisters of a respondent. We note that variables in this class based on recall information were all enquiring about performance in public examinations. It is not possible to say whether their slightly lower reliability is primarily due to their formal characteristics of 'recall' or their substantive content (examination performance), or to some combination of both.

The estimates of true score reliability for the factual questions about others suggest that the school leavers are able to answer questions about the education of both parents and about their father's occupation equally reliability.

[Table 4 about here]

All questions in the model based on the question type of 'non-factual/ reason questions about oneself' refer to substantive issues of why the leaver chose to pursue further education after leaving school. There appears to be no strong differentiation in reliability between the questions in this model. The MMMQ estimates suggest that the most salient distinction may not be one of substantive content but rather one of the specificity of the question. Questions directed at an individuals' specific objective appear to be more reliable than those directed at a more general objective in pursuing further education. This conclusion must, however, be very tentative, because it is not supported by the correlational estimate of reliability for the ZR data and it did not prove possible to fit an identified MMMQ model to this (ZR) set of data.

The questions in the model based on the 'non-factual/perceptual questions about oneself' all related to a substantive content dealing with how well the school leavers' teachers knew them in their last year at school. The main finding for this model is that evaluations relating to teacher advice on study are much less reliable than those on advice relating to future education/training, future jobs or personal matters. The percentages of respondents in the two data sets who said that some of their teachers knew them well enough to give advice on the different

issues asked about are give in Table 4. The main point to be taken from this table is that the aggregate percentages are relatively constant for the same question across both questionnaires for the two samples. However, the results for the MMMQ model indicated relatively low true score reliability for all of these questions, thus the overall picture of the aggregate percentages masks substantial variation in response within individuals across the two occasions. This strong variation within individuals across occasions means that comparisons across questions are of secondary importance. In Table 4 there seems to be no discernable pattern between the aggregate percentages and the level of reliability across questions, thus reliability does not seem to covary with the percentage marking one or other of the binary response categories. Perhaps the most useful conclusion to draw for this particular model and data, is that there appears to be a great deal of scope for falacious interpretation if aggregate statistics are used to try and make inferences about individuals qua individuals.

1.4 Discussion

The findings reported above point to one main and one subsidiary conclusion from this study. The main conclusion is that the substantive content of questions asked in postal surveys of school leavers does not appear to markedly influence the reliability of responses. Of far greater importance in determining the reliability of responses are the formal characteristics of the questions. An earlier report on this research (Cuttance, 1981b) has shown that the typology put forward in section 1.1 as a classification of question/variable types may have a significant potential to inform survey practice. Further, it provides a framework for the specification of covariance structure models from which components of reliability can be estimated. The general conclusion must, however, be tempered by the knowledge that the substantive content of questions also varied across models. Thus some of the variation attributed to the formal characteristics of questions could also be attributed to their substantive content. Data just collected on a wider range of variables will allow this qualification to be tested within a more complex factorial design.

The subsidiary conclusion to be drawn is that, estimates of reliability based on correlation coefficients under the assumption that the measures are parallel may be biased either upward or downward. The direction and extent of such bias is dependent upon the behaviour of systematic extraneous influences on the observed responses.

Part II of this paper now follows. It offers a salient example of why the estimation of reliability coefficients is of both methodological and substantive importance for the problem of answering questions about the relationships between theoretical constructs, when the data available are for fallible observations on the manifest counterparts of these constructs in the empirical world.

Part II: The Effects of Response Error on Estimates of Correlation and Regression Coefficients

2.1 Accounting for Measurement Error in the Analysis of Data

In the context of the analysis of Part I the theoretical constructs of central interest were the 'true scores' assumed to underly the observations on the manifest variables. The analyses there indicated that the observations on the manifest variables can be thought of as consisting of at least three components, due to: the true score, a method/occasion effect and, a random error term.

In most multivariate analyses of social science data the researcher is interested in the relationships between particular theoretical constructs. These constructs may be represented by the 'true scores' for their directly observed manifest variables, or they may be represented by a weighted combination of manifest variables. The former type of construct derives its use from Classical Test Theory and can be thought of as the indirect measurement of a theoretical construct through the replicated observation of parallel or tau-equivalent items. The latter type of construct is based on the notion that the common variation in a set of manifest variables or 'indicators' represents the essence of the construct. An example of the former may be that of 'age'

as the theoretical construct and the reply to a question 'How old were you at your last birthday?' as the observation on one manifestation of that construct. In this case it is assumed that the respondent has a 'true' age but that the observed response may be influenced by other factors also. Replicated observation on this question captures the true age of the respondent. An example of a multiple indicator construct is that of 'social status', which may be indirectly measured by observations on the manifest variables of education, income and, say, wealth. Here the researcher will hope to obtain a measure of the theoretical construct as a weighted combination of the scores observed for the manifest variables. The appropriate weighting is determined by each indicators' contribution to the common variance between them. In the context of a path diagram, the manifest variables are referred to as indicators of the theoretical construct, which is represented as a 'latent' variable. Relationships between the latent variables are referred to as the structural model and those between the latent and manifest variables represent the measurement models for the theoretical variables. In such a model (see Figure 5) there are thus no direct 'causal' relationships between the manifest variables.

[Figure 5 about here]

In order to analyse data and directly take account of response error it is necessary to use one of the structural equation computer programs (LISREL-Joreskog and Sorbom, 1981; EQS - Bentler and Weeks, 1980; COSAN-McDonald, 1979) ,which provide consistent unbiased estimates of the parameters. In these programs, all relationships in the model are estimated simultaneously. However, a second best solution is also possible if the researcher does not have access to such programs. This involes a two stage procedure of which the first stage is to estimate the measurement models and compute weighted scores for each of the latent variables. The second stage then consists of estimating the relationships between the latent variables. This solution is not as efficient, nor does it allow for the same complexity of relationships as that available in the full system approach. If the response error in the observed variables is entirely composed of random errors of measurement,then these two procedures are formally equivalent. However,

if the response error contains a component of systematic variation then the correlation coefficients based on the observed variables will be biased, in addition to being attenuated.Under such conditions the two stage procedure will lead to biased estimates in correlation and regression analyses. Both approaches may, however, to be preferred to one in which the observed variables are treated as completely reliable measures of the theoretical constructs of interest, as is the case, for example, in most applications of ordinary multiple regression analysis.

In summary, we can distinguish three levels of sophistication in dealing with fallible data when undertaking multivariate analyses:

(1) treat the data as if it were perfectly reliable or,

(2) use a two stage procedure to correct each of the measures for the attenuation effects of unreliability and, then perform the multivariate analysis or,

(3) use a full system (full information) approach which simultaneously corrects for the attenuation due to unreliability and estimates the relationships between the theoretical variables of interest.

The effect of less than perfect reliability on bivariate correlation is described as attenuation of the magnitude of the coefficient , ie. it leads to a reduction in the magnitude of the correlation between the observed variables compared to that which actually exists between their true scores. Adjusting the correlations between observed variables for their respective reliabilities is known as disattenuating the 'observed correlations'.

The usual formula employed in disattenuating a Pearson product moment correlation when the response error is entirely due to random measurement error is:

[Equation 1 about here]

(Lord and Novick, 1968).

Equation 1 involves division by the product of the reliability coefficients for each variable. Thus, it is important that the estimates of reliability should not be biased downward if we are not to overcorrect for the effects of unreliability. Many commonly employed coefficients of composite (latent variable) reliability, eg. Cronbach's alpha, and a range of coefficients based on the classical test theory model (cf.Bentler and Woodward, 1980), are lower bound estimates of the true reliability of the composite. Using these lower bound estimates to disattenuate correlations between latent variables will lead to over correction and inflated disattenuated correlations. General practice is to recommend that the highest estimate of reliability obtained from several different estimators be employed in the disattenuation formula. However, a more appropriate recommendation would be to 'model the response error' and check that the measurement model provides a suitable fit to the data before using the corresponding reliability coefficient. This generally requires the use of the most sophisticated of the approaches noted above.

2.2 Effects of Response Error on Estimate of the Correlation Coefficient

Table 5 presents estimated attenuated and disattenuated correlation coefficients for a subset of the variables in this study. In disattenuating these coefficients I have used the estimates of true score reliability from the MMMQ models presented in Table 3. The variables have been listed in the table in descending order of reliability. There is close agreement between the attenuated and disattenuated correlation coefficients for pairs of variables with high reliability but, as reliability decreases, the degree of difference between the two types of coefficient increases. Thus, the greatest degree of difference between attenuated and disattenuated correlations is to be found in the bottom right of the table. The margin of difference in the magnitudes of the coefficient depends on both the value of the attenuated correlation and the level of reliability of the variables.

[Table 5 about here]

From the differences in size between the attenuated and disattenuated correlation coefficients it is evident that multivariate analyses of correlations (for example, factor and regression analysis) could be quite misleading if based on attenuated coefficients. Only in cases where we can be relatively sure that the measures are of a reliability close to unity can we base the analyses on attenuated coefficients.

The degree of reliability of a variable imposes a constraint upon the maximum correlation which can be observed between itself and another variable. For example, the conventional wisdom suggests that an attitudinal variable with reliability as low as 0.60 is suitable for use in multivariate analyses. The maximum correlation observable between two variables of such reliability is itself 0.60. In conventional terms this magnitude of correlation between two attitudinal variables would be considered to indicate something of substantive importance in the relationship between these manifest variables. However, the substantive importance of the relationship would be seriously underestimated; for to observe a correlation of 0.60 when the reliability of the measures was also 0.60 would require that the magnitude of the correlation between the theoretical variables was unity. Looked at another way, the two observed variables would have only 36 per cent of their individual variances in common, whereas their true scores would have all (100 per cent) of their variances in common.

2.3 Effects of Response Error on Regression Equations

A regression equation for attenuated and disattenuated coefficients is presented in Table 6. This equation was estimated by the full system approach outlined previously using the LISREL IV computer program. The model estimated did not allow for covariation between the error terms in the measurement models and this has resulted in implied reliabilities for the observed variables which are somewhat higher than those estimated in Part I by the MMMQ models. When I attempted to estimate a regression equation based on a correlation matrix disattenuated by the

MMMQ estimates of reliability, it was found that the matrix could not be inverted due to redundancy among the variables. That is, when full correction for attenuation was applied to the variables, it was found that the intercorrelations between one (or more) of the variables and the other variables in the matrix was an exact linear sum of the intercorrelations between that variable and other variables in the matrix. The only solution to this computational problem would be to delete one or more variables from the matrix. Rather than report results for different attenuated and disattenuated equations, I have chosen to report the findings for the model disattenuated as in Figure 5. The higher estimates of reliability implied by this model, compared to the MMMQ estimates, are due to the error terms between indicators of a given true score being specified as uncorrelated. This is necessary in order to identify the model and hence estimate the structural relationships between the dependent and independent latent variables. These estimates are the structural model equivalent of an ordinary multiple regression between the disattenuated observed variables. I expect these estimates to be biased, but less so that those based on an attenuated regression equation. Further, the model employed is a fair representation of the degree of disattenuation undertaken in many of the structural equation models presented in the literature.

[Figure 5 about here]

[Table 6 about here]

The results in Table 6 show that the estimation and interpretation of coefficients in regression equations are highly sensitive to attenuation effects due to response error. The standardised beta coefficients can be grossly underestimated by attenuated equations and, perhaps more significantly, the direction of the relationship between the dependent and independent variables can be incorrect. When the equation contains variables of varying reliability, all coefficients appear to be subject to error, not only those of lower reliability. The disattenuated equation accounts for a much higher proportion of the variance in the dependent variable, indicating that the relationship between the theoretical constructs (as represented by the true scores) is much

stronger than would appear from the observed manifest variables. This point is further demonstrated by the large number of changes in the level of significance of the beta coefficients. In some cases, attenuated coefficients which are not significantly different from zero at the 5 per cent level are found to be significantly different from zero at the 1 per cent level in the disattenuated equation. This is primarily due to the fact that changes in the standard error of the estimate are proportionately much less than changes in the magnitude of the beta coefficients between the attenuated and disattenuated equations.

2.4 Discussion

The findings of Part II amply demonstrate the reasons of why models of the type presented in Part I are necessary in the analysis of multivariate data by correlational methods.

On theoretical grounds we expect the model of disattenuation employed in Part II to yeild less biased estimates of the true relationships between theoretical constructs than that which fails to apply any correction at all for response error. The degree of disattenuation attempted here shows that interpretation based on attenuated regression equations may be seriously misleading. Further, the computational problem experienced in attempting a full disattenution of the equation in this analysis suggests that the interpretation of the data could be even more parsimoneous than that based on the disattenuated equation presented in here.

The general conclusion to be derived from the findings presented above is that the accepted practice of treating survey data as infallible, on the assumption that it is probably nearly so, is not warranted without a major exploratory investigation of the reliability of the data. Findings based on analyses which fail to take account of response error may not only underestimate the strength of relationships between variables, they may dismiss relationships as trivial when those between the theoretical constructs they represent may be substantial in

both statistical and theoretical terms. Perhaps the greatest errors in theoretical interpretation will, however, be made in assessing the direction of relationships, particularly when this error is likely to interact with others.

All of these problems can be handled simultaneously by recent developments in covariance structure and structural equation modelling. For an overview of these developments see Bentler (1979), Bielby and Hauser (1978) and, Carmines and McIver (1981). My own 1981 paper (see references) attempts to provide a basic introduction to the LISREL version of such models as developed by Joreskog and Sorbom (1981). It is time that educational researchers and survey designers made themselves more familiar with the capabilities of these approaches. I hope I have shown one substantial reason why such training is not only worth the effort but methodologically important for the future of educational research.

References and Relevant Papers

Alwin, Duane F. and David J. Jackson (1980): 'Measurement models for response errors in surveys: issues and applications'. Sociological Methodology 1980 ,68-119.

Bentler,P.M.(1980): 'Multivariate analysis with latent variables: causal modelling'. Annual Review of Psychology , 31 ,419-456.

Bentler, P.M and D.G. Weeks (1980) : 'Linear structural equations with latent variables'. Psychometrica ,45 ,298-308.

Bentler, P.M. and J. Arthur Woodward (1980): 'Inequalities among lower bounds to reliability: with applications to test construction and factor analysis'. Psychometrica, 45, 249-267.

Bielby,William T. and Robert M. Hauser (1977): 'Structural equation models'. In A.Inkeles,J.Coleman and N.Smelser (Eds), Annual Review of Sociology ,3. Annual Reviews:Palo Alto,California.

Bielby, William T., Robert M Hauser and David L. Featherman (1977): 'Response errors of non-black males in models of the stratification process'. In D. S. Aigner and A. S. Goldberger (Eds), Latent Variables In Socioeconomic Models. North-Holland Publishing Company: Amsterdam.

Blalock, H.M. (1969): 'Multiple indicators and the causal approach to measurement error'. American Journal of Sociology ,75 , 264-272.

Campbell,D T. and D. W.Fiske(1959): 'Convergent and discriminant validation by the multitrait-multimethod matrix'. Psychological Bulletin , 56,81-105.

Carmines, Edward G. and John P. McIver (1981): 'Analysing models with unobserved variables'. In George F. Bohrnstedt and Edgar F. Borgatta (Eds), Social Measurement.Sage:California.

Carmines, Edward G. and Richard A. Zeller (1979): Reliability and Validity Assessment. Sage University Papers: Quantitative Applications in the Social Sciences. Sage Publications: Beverley Hills, California.

Cuttance, Peter F. (1981a): 'Covariance structure and structural equation modelling in educational research: A conceptual overview of LISREL modelling'. Paper presented at a Social Research Association Symposium on Structural Equation Methods in the Social Sciences with Particular Reference to LISREL,University of London Institute of Education,Sept.

Cuttance,Peter F. (1981b) : 'An application of LISREL modelling to the estimation of reliability and method effects in a sociological survey'.Working Paper No.8102,Centre for Educational Sociology, University of Edinburgh.

Duncan, O.D. (1966): 'Path analysis: sociological examples'. American Journal of Sociology ,72, 1-16.

Ecob, Russell and Harvey Goldstein (1981): 'Instrumental variable methods for the estimation of test score reliability'. Research paper, Institute of Education, University of London.

Greene,Vernon L. and Edward G. Carmines (1980): 'Assessing the reliability of linear composites'. Sociological Methodology 1980 , 160-175.

Heise, D. and G.W. Bohrnstedt (1970): 'Validity, invalidity and reliability'. In E.F. Borgatta and G.W. Bohrnstedt (Eds), Sociological Methodology. San Fransisco: Jossey-Bass.

Jackson, David J. and Edgar F. Borgatta(Eds) (1981): Factor Analysis and Measurement in Sociological Research.California:Sage Publications Inc.

Jones, C.L. and A.F. McPherson (1972): 'Implications of non-response to postal surveys for the development of nationally based data on flows out of educational systems'. Scottish Educational Studies, 4 (1), 28-38.

Joreskog, K G.(1971): 'Statistical analysis of sets of congeneric tests'. Psychometrica , 36,109-133.

Joreskog, K.G. and D. Sorbom (1976): 'Statistical models and methods for test-retest situations'. In D.N.M. Gruijter, L.J.Th. van der Kamp and H.F. Crombag (Eds), Advances in Psychological and Educational

Measurement ,New York: Wiley. 135-157.

Joreskog, K.G. and D. Sorbom (1977): 'Statistical models and methods for analysis of longitudinal data'. In D.J. Aigner and A.S. Goldberger (Eds), Latent Variables in Socioeconomic-Models. Amsterdam: North-Holland Publishing Co. 285-325.

Joreskog, K.G. and D. Sorbom (1978) : LISREL IV User's Guide Chicago: National Educational Resources

Joreskog, K.G. and D. Sorbom (1979): Advances in Factor Analysis and Structural Equation Models. Abt Books: Massachusetts.

Joreskog, K.G. and D.Sorbom (1981): LISREL V Users Guide. Chicago:International Educational Services.

Kalton, Graham and Howard Schuman (1981): 'The effect of the question on survey responses: a review'. Proceedings of the American Statistical Association Conference 1981 ,forthcoming.

Lord, F.N. and M.R. Novick (1968): Statistical Theories of Mental Test Scores. Reading: Addison-Wesley Publishing Company.

Mason, William M. et al (1976): 'Models of response error in student reports of parental socioeconomic characteristics'. In Sewell, W.H., Hauser, R.M., and Featherman D. L. (Eds), Schooling and Achievement in American Society. New York: Academic Press.

McDonald,Roderick P. (1979): 'The structural analysis of multivariate data: a sketch of a general theory'. Multivariate Behavioral Research ,14,21-38.

Novick, M.R. and C. Lewis (1967):, 'Coefficient alpha and the reliability of composite measurements'. Psychometrica, 32, 1-13.

Raffe, D., J. Lamb, F. Cassels, E. Cope, J. Gray, A. McPherson and L. Wylie (1977): Collaborative Research Dictionary 1977. Centre for Educational Sociology, University of Edinburgh.

Zeller, Richard D. and Edward G. Carmines (1980): Measurement in the Social Sciences. Cambridge University Press: N J.